"What is Essential for Virtual Reality to Meet Military Performance Goals" Performance Measurement in VR

Lt. Jim Patrey[1], Robert Breaux, Andrew Mead & Elizabeth Sheldon
Virtual Environment Training Technology (VETT) project
Naval Air Warfare Center Training Systems Division (NAWCTSD)
12350 Research Parkway
Orlando, FL 32826-3275
U.S.A.

One of the unique attributes and potentially greatest assets of virtual environments is the unique ability to comprehensively measure human performance. In the real environment, measuring human behaviors is usually, though not always, feasible and typically extremely effort intensive and cost-prohibitive. Similarly, there is substantial environmental variability that can have pervasive effects on human performance, but is beyond any feasible, economic data capture. Virtual environments instill the capability for comprehensively monitoring both user inputs and interactions and the environment (as well as control the virtual environment and thereby eliminating confounding variables with precision beyond that of real environment lab research).

Monitoring and measuring human behavior in this fashion provides three invaluable elements. Firstly, it furnishes a valuable research tool for the development of outcome measures for performing research. Secondly, performance measurement has training value for assessment and evaluation. The derivation of accurate performance measures can enable improved proficiency and reduced training time when implemented in a training curriculum. Finally, the development of performance measures can facilitate the development of intelligent tutoring systems and thereby cost-effective, stand-alone training systems. Measuring human performance can be of great use in the facilitation and maximization of training.

Performance measurement involves three distinct processes: Identification, Monitoring, & Evaluation. *Identification* is the determination of the significant measures of performance for a given task. This is typically accomplished via cognitive task analysis and intense subject-matter expert (SME) interviews and observation and/or statistical analytic techniques occurring after the observation of real world performance. These approaches are the two traditional approaches to performance measure development.

The advent of virtual environments has fostered the development of two new approaches to the development of performance measures - cognitive model driven and data-driven approaches. Cognitive models enable a new method of performance measurement. Through traditional approaches (such as SME interviews) a cognitive model can be developed for a given task (in truth, a cognitive task analysis is a variant of a cognitive model, typically represented in GOMS format). There are a host of cognitive modeling approaches (discussed in detail in Pew & Mavor, 1996), but they all generally afford identification of cognitive variables not easily discernable through traditional approaches. However, the usefulness of such models for performance measurement is dependent on the accuracy of the model and the development of cognitive models can be resource-intensive, particularly for complex tasks.

Data-driven approaches are also afforded by virtual environments. The ability to thoroughly monitor and record all actions and interactions in a virtual environment enables data mining approaches to provide value to the determination of performance measures. There are numerous data-driven techniques for mining data (such as neural networks, genetic algorithms, evolutionary computing, etc.), but it is fuzzy sets theory, or fuzzy logic, which may hold the most promise for identifying crucial aspects of human performance. Unlike other approaches, fuzzy logic preserves the semantic value of the input variables. Output from fuzzy models meaningfully represents human behavior and can be directly applied to performance measure development (Cowden, Burns, Casey, & Patrey, 2000).

It is likely that all of these approaches should be integrated to fully profit from virtual environments for the identification of performance measurement. Ideally, we will someday be able to place a SME in a VE to perform a task and have hybrid models (of both top-down cognitive models and bottom-up data-driven models) monitor the virtual world and generate performance models that produce measures of performance.

[1] For correspondence with author: patreyje@navair.navy.mil

Paper presented at the RTO HFM Workshop on "What Is Essential for Virtual Reality Systems to Meet Military Human Performance Goals?", held in The Hague, The Netherlands, 13-15 April 2000, and published in RTO MP-058.

VE-based performance measures cannot be developed without *Monitoring* the virtual environment. Accomplishing this requires monitoring behaviors and their consequences within the VE. Behaviors include active behaviors such as control inputs and verbal commands as well as passive behaviors such gaze surveys. The consequences of these actions include movement through the VE and interactions with and within the VE resultant from user behaviors. The principal behaviors and consequences must accurately represented, inherently measurable, and recorded for the effective use of VR for performance measurement.

Implicit in this is the indispensability of adequate modeling of the VE. All salient cues must be represented with suitable fidelity within the VE for the performance measures reaped to represent real world performance. This may be the greatest challenge for the practical use of VE for performance measurement. It generally behooves VE developers to minimize the fidelity in order to minimize processing demands and cost. The level of fidelity should be mapped to the task fidelity requirements so that 'training' fidelity, the level of fidelity required to meet training requirements, can be attained. The role of the SME cannot be underestimated in fulfilling this balance between minimal fidelity and requirements. Achieving this necessitates thorough front-end analysis prior to significant investment in development of the VE.

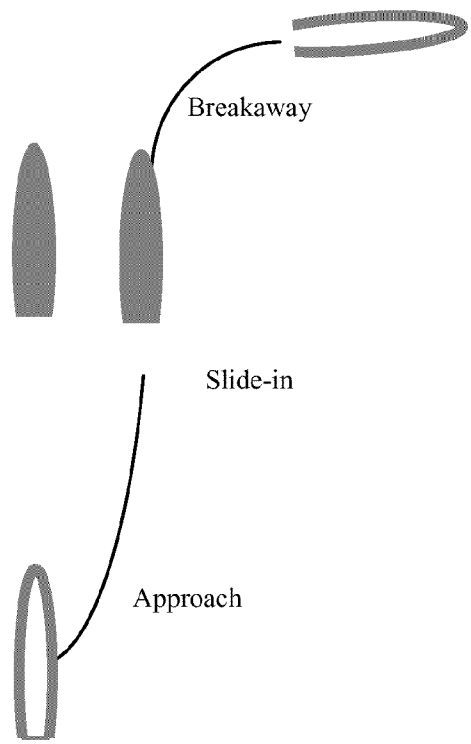
Finally, the effective use of VE in performance measurement should also provide performance *Evaluation*. Beyond identifying and monitoring performance measures is the need to discriminate good/expert performance from bad/novice performance. This is most meaningful for VE in the context of developing intelligent tutoring systems (ITS), but also permits structured, empirically based, objective feedback in any circumstance.

Derivation of evaluatory measures of performance (MOPs) is generally accomplished through methods similar to identifying performance measures. Traditional methods include SME ratings of performance (typically gathered through observation of another's performance) and statistical analysis. Cognitive model and data driven approaches also hold promise for evaluating performance (particularly in contrasting novices and experts), but they have not been applied as extensively in this domain.

Traditional performance measure development for virtual underway replenishment

An immersive virtual environment has been developed for underway replenishment (UNREP) with a U.S. Navy Cruiser (see Davidson, 1997 and Martin et al., 1998 for more information on the virtual UNREP). An UNREP involves the transfer of fuel, stores, ammunition, and people from one vessel to another while underway. It is comprised of four distinct phases (see Figure 1): 1) *Approach* - from awaiting station to bow-stern crossing (overtake oiler & attain lateral separation), 2) *Slide-in* - transition from approach to alongside (match velocity), 3) *Alongside* - stationkeeping (maintaining proper lateral separation and matched velocity), & 4) *Breakaway* - separation of own ship from oiler.

Figure 1. Depiction of the phases of Underway Replenishment.



The ship is controlled by the Conning Officer via verbal commands to a virtual helmsman. The verbal commands are broken down into two main types: control commands and requests for information. Control commands include engine commands such as all stop, all back, all ahead, indicate knots, increase turns, & decrease turns and rudder commands such as rudder amidships, steer course, left rudder, & right rudder. The Conning Officer can also make "requests for information" regarding rudder angle, relative bearing, true bearing, heading, speed, & range. These shiphandling behaviors provide a solid foundation upon which to develop MOPs.

Iterative inputs from SMEs also identified ship dynamic features indicative of good performance. These parameters vary depending upon the phase stage (approach, slide-in, alongside, or breakaway), but generally include relative positional data (vertical separation, lateral separation, & bearing) and relative velocity. The following depicts the statistical analyses conducted in pursuit of MOP identification.

Method

Subjects

Twenty-six (26) male Navy personnel (students & instructors) of the Surface Warfare Officer's School (SWOS) participated as subjects. Due to technical errors in the VE data collection process, data from eight (8) subjects, were not included in the analysis. The level of duties represented in the sample were Ensign (ENS, n = 6), Division Officer (DIVO, n = 4), Department Head (DH, n = 3), and Commanding/Executive Officer (CO/XO, n = 5). Further description of the subject demographics can be found in Martin, Sheldon, Kass, Mead, Jones, & Breaux (1998).

Apparatus

The VE testbed was comprised of the following hardware: Dual Processor Octane R 10000 Processors, MXI Graphics, Octane Channel Option, and Indigo^2 Impact R 10000 IDS by *Silicon Graphics, Inc.* Subjects used a VR4 Head Mounted Display (HMD) by *Virtual Research,* and IS600 Inertial Tracker by *Intersence* to view the graphics. The commercial software components were dVise by *Division* and Vega Marine by *Paradigm.* Further specifications can be found in Davidson (1996,1997a, 1997b).

Questionnaires

Subjects were administered six questionnaires: Pre-Questionnaire, Demographics Questionnaire, Pre-Exposure Symptom Checklist, Scenario Review, Post-Exposure Symptom Checklist, and Debrief. The Pre-Questionnaire and Demographics Questionnaire were completed prior to the experimental session. The Pre-Questionnaire solicited comments regarding the critical points of an UNREP, UNREP performance measurements, typical UNREP strategy, and a diagram of the UNREP outlined in the strategy. The Demographics Questionnaire gathered background information on shiphandling, UNREP, and VE experience. The Scenario Review was administered between the performance of the two VE UNREPs to obtain the subject's appraisal of the first UNREP and planned strategy modifications for the second pass. The Debrief was given after the performance of the second UNREP to acquire a comparison of the two UNREPs and usability comments. The results of the usability comments are described in Martin et al. (1998). The Pre- and Post- Exposure Symptom Checklists, an adaptation of the Simulator Sickness Questionnaire (SSQ, Kennedy et al., 1993; Lane & Kennedy, 1988), were used to examine the occurrence of simulator side effects and will be described in a future report.

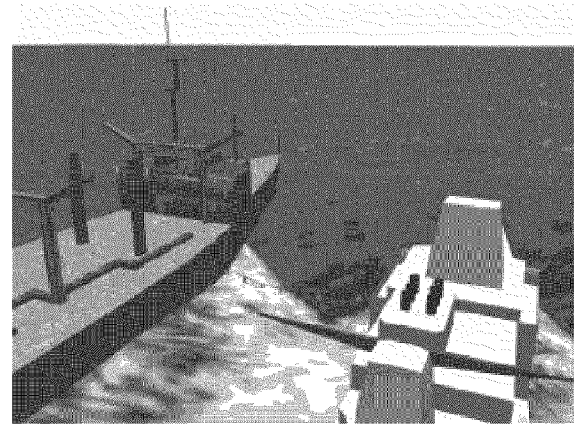
VE UNREP Scenario

The scenario task was to execute an UNREP from the port bridgewing of a guided missile cruiser (CG) and conn the ship alongside a supply ship, maintain the alongside position (at 120 feet lateral separation) for two minutes, and breakaway from the supply ship (see Figure 2 for alongside view). At the scenario start, ownship was positioned 1000 yards directly behind the supply ship, and both ships were traveling on a heading of 130° at a speed of 15 knots (the UNREP course and speed).

Procedure

Subjects received a review sheet (an informative briefing of the VE ship's characteristics, general reminders regarding hydrodynamic effects, and rules of thumb applicable to UNREP) to study prior to the experiment.

Figure 2. Virtual Underway Replenishment.



The session began with the subject's review of written instructions describing the task and pictures of the location of the supply ship's UNREP station displayed on a PC monitor. The subjects were instructed to issue commands and requests for information as in the real world. These commands and information requests were input to the simulator by an experimenter via keyboard strokes. Replies to commands were made by a pre-recorded speech system, and replies to requests for information were provided verbally by the experimenter. The subjects completed two UNREPs and were given a brief rest period between the UNREPs in which they completed the Scenario Review. It took approximately 1.5 hours to complete the entire experimental session. The first UNREP was considered a practice trial enabling subjects to adapt to the VE. The second UNREP was used for all subsequent analyses.

Following UNREP performance, SMEs were solicited to rate UNREPs presented as plot tracks. Six experienced Surface Warfare Officers rated performance by evaluating a printed track of each subject's UNREP performance. Each track was assigned a rating of 0 to 100. One rater who demonstrated poor internal consistency and poorly correlated with the group was dropped. The mean inter-rater correlation of the remaining five raters = .68; ranging from .56 to .78. The ratings from the five remaining raters were averaged to derive a final performance rating for each UNREP.

Results

The experience level of the sample was diverse, ranging from ensign to commanding officer with a median of 8 years shiphandling experience. The median number of

deployments completed was 4 and the median elapsed time since the last deployment was 3 years. A typical UNREP has an extended duration. Depending on the type of ship, an UNREP can last as long as12 hours (though 1 to 3 hours is more typical), therefore several officers assume the conn during a single evolution. The subject's UNREP experience included completion of a median of 17 approaches, a median of 22 alongsides, and completion of a median of 10 breakaways.
Pursuit of good performance measures began with evaluation of requests for information, engine & rudder commands, & ship dynamic characteristics.

Requests for information (RFI)
Difference comparisons between novice ensigns (no shiphandling experience; n=6) and experienced shiphandlers (n=12) were made for RFI (rudder angle, relative bearing, true bearing, heading, velocity, & range). Novice shiphandlers made significantly more requests for velocity (Novices = 6.3, Experts = 3.0; One-way ANOVA, $F=2.40$, $p<.05$) and relative bearing (Novices = 11.2, Experts = 3.3; One-way ANOVA, $F=6.99$, $p<.01$). These differences are consistent with rules of thumb that novices are taught to judge relative positions; experienced shiphandlers rely instead on "seaman's eye" (Crenshaw, 1965) and rarely use these rules and therefore don't make the same RFIs.

In order to determine whether any RFIs were predictive of performance, a linear regression model of SME ratings from RFI was conducted and produced an $R=.48$ ($F=0.59$, ns). No individual RFIs were statistically significant in this model. This suggests that RFIs are not effective measures of performance, though they do appear to be indicative of experience.

Engine & Rudder commands
One-way ANOVAs were conducted comparing novice and expert shiphandlers on their cumulative use of shiphandling commands; none of the comparisons on these engines and rudder commands were statistically significant. Furthermore, a linear regression predicting SME ratings from these shiphandling commands was also not significant ($R=.47$, $F=0.91$, ns).

Ship dynamics
Candidate measures of ship dynamics as characteristic performance measures were gathered from SME interviews and prior shiphandling dynamics analyses (Martin et al., 1998, Patrey et al., 2000). The most meaningful single relative position, based upon these prior analyses, is within the transitional slide-in phase; in particular, the ship dynamic characteristics (lateral separation, bearing, velocity, & acceleration) at approximately 100 feet astern of the stationkeeping position appears to be the single most distinguishing point. Additionally, measures from the alongside phase for minimum lateral separation (LS), maximum LS, root mean square (RMS) LS, & RMS vertical separation (VS) were included as potentially significant measures.

A linear regression predicting SME ratings from these ship dynamic characteristics was highly significant ($R=.98$, $F=15.76$, $p<.001$). In order to create a more parsimonious model, a backward elimination linear regression predicting SME ratings from this host of variables reduced the model to velocity, relative bearing, LS, maximum LS, & RMS LS ($R=.92$, $F=12.99$, $p<.001$).

Discussion
Performance measures were successfully identified for virtual UNREP using a traditional approach of identification. Indices of relative position (LS, RMS LS, & maximum LS), relative velocity, and relative bearing significantly predict SME evaluation of performance. Iterative development of the VE coupled with feedback and inputs from SMEs and data analysts enabled the monitoring of salient measures of performance (such as ship dynamics). Furthermore, this has provided a basis for empirically driven performance evaluation.

This clearly demonstrates the functionality of using VE as a tool for deriving performance measures for a real world task. Collecting this quality of data in the real world is a daunting task (though efforts are underway to accomplish this to validate matching between real and virtual UNREPs). While possible to collect this data in the real world, it is difficult and uneconomical to do so, particularly when VE affords an alternative, potentially more effective, method for accomplishing this.

While this particular performance measure derivation effort was primarily driven by a traditional approach to knowledge extraction, virtual data was manually processed with standard statistical methods to glean performance measures that were not wholly apparent from SME interviews. This highlights the need, for at least some types of task, such as those heavily perceptual in nature and not easily verbalized, for additional methods of knowledge elicitation.

Data and cognitive model driven approaches were discussed as potential methods of facilitating and streamlining the knowledge acquisitions process. Currently, both approaches are being investigated for virtual UNREP. Fuzzy logic, as a data driven approach, and COGNET (Cognitive Network of Tasks, Chi Systems Inc.), as a cognitive modeling approach, are the platforms of choice for virtual UNREP and will provide some guidance as to the value in using these powerful tools for performance measure extraction.

This is likely where one of VE's great potential can be realized — as effectual and inexpensive generators of performance indicators, monitors of performance, and ultimately providers of performance evaluation. As these data mining cognitive modeling tools continue to develop, their integration within VE, particularly VE training systems, may prove to be the cornerstone in the revolution in training.

References

Cowden, A.C., Burns, J. J., & Patrey, J. (2000). *Data-driven knowledge engineering.* To be presented at the Interservice/Industry Training, Simulation, & Education Conference.

Crenshaw, R.S., Jr., CAPT, U.S. Navy (Retired). (1975). *Naval Shiphandling.* Annapolis, Maryland: Naval Institute Press.

Davidson, S. (1996). *Software Design of a Virtual Environment Training Technology Testbed and Virtual Electronic Systems Trainer* (Technical Report 96-002). Orlando, FL: Naval Air Warfare Center Training Systems Division.

Davidson, S. (1997a). *Design of an Open Water Shiphandling Software Testbed* (Technical Report 97-007). Orlando, FL: Naval Air Warfare Center Training Systems Division.

Davidson, S. (1997b). Development of a virtual environment software testbed using commercial off the shelf software components. *In Proceedings to NATO VR Conference*, (December, 1997), Orlando, FL.

Kennedy, R.S., Lane, N.E., Burbaum, K.S., & Lilienthal, M.G. (1993). A simulator sickness questionnaire (SSQ): A new method for quantifying simulator sickness. *International Journal of Aviation Psychology, 3 (3),* 203-220.

Lane, N.E. & Kennedy, R.S. (1988). *A New Method for Quantifying Simulator Sickness: Development and Application of the Simulator Sickness Questionnaire (SSQ).* (Technical Report EOTR 88-7). Orlando, FL: Essex Corporation.

Martin, M.K., Sheldon, E., Kass, S., Mead, A., Jones, S., & Breaux, R. (1998). Using a virtual environment to elicit shiphandling knowledge. *Proceedings to '98 I/ITSEC Conference*, Orlando, FL, December 1998.

Patrey, J., Sheldon, E. M., Breaux, R.B., & Mead, A. M. (2000). Quantifying performance of a dynamic shiphandling perceptual-action task. Technical Report in preparation.

Pew, R. W., Mavor, A. S., *Modeling human and organizational behavior: application to military simulations,* National Academy Press, Washington, D.C. 1998.

This page has been deliberately left blank

Page intentionnellement blanche